

Runaway Events Dominate the Heavy Tail of Citation Distributions

Michael Golosovsky and Sorin Solomon ^a

The Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

Received: date / Revised version: date

Abstract. Statistical distributions with heavy tails are ubiquitous in natural and social phenomena. Since the entries in heavy tail have unproportional significance, the knowledge of its exact shape is very important. Citations of scientific papers form one of the best-known heavy tail distributions. Even in this case there is a considerable debate whether citation distribution follows the log-normal or power-law fit. The goal of our study is to solve this debate by measuring citation distribution for a very large and homogeneous data. We measured citation distribution for 418,438 Physics papers published in 1980-1989 and cited by 2008. While the log-normal fit deviates too strong from the data, the discrete power-law function with the exponent $\gamma = 3.15$ does better and fits 99.955% of the data. However, the extreme tail of the distribution deviates upward even from the power-law fit and exhibits a dramatic "runaway" behavior. The onset of the runaway regime is revealed macroscopically as the paper garners 1000-1500 citations, however the microscopic measurements of autocorrelation in citation rates are able to predict this behavior in advance.

PACS. 01.75.+m Science and society, – 02.50.Ga Markov processes, – 89.75.Fb Structures and organization in complex systems, – 89.75.Da Systems obeying scaling laws.

1 Introduction

Dynamic statistical distributions found in nature frequently exhibit heavy tails that are usually approximated using log-normal or power-law functions [1]. However, there can be a few individual values in the extreme end of the distribution that exceed by far both these fitting functions. We advance the hypothesis that these "runaways" can eventually capture the entire system. Statistical distribution of citations of scientific papers offers a good example of such dynamics.

Intensive studies of citation statistics were triggered by the seminal work of de Solla Price [2] who showed that the statistical distribution of citations of scientific papers has heavy tail that can be approximated by a power-law function. De Solla Price proposed a microscopic "cumulative advantage" mechanism [3] that in the long time limit generates such power-law distribution. Following his line of thought, subsequent studies of citation statistics [4, 5, 6, 7, 8, 9] used power-law probability distributions to fit their data. Later on, it turned out that stretched exponential [10, 11, 12] or log-normal [13, 14, 15, 16] functions fit the measured citation distributions equally well. Ref. [17] showed that the choice between the power-law and the log-normal fits is extremely difficult. Indeed, the most important difference between the two is that the power-law decays slower than the log-normal, in other words it

has heavier tail. To probe this tail one has to measure a very large data set in order to achieve enough entries in the tail. Previous studies of citations of scientific papers used either small [14, 15] or inhomogeneous [13] data sets that were insufficient to distinguish between the log-normal and power-law fits.

The initial goal of our study was to find out which function fits better the citation distribution in the long time limit: the power-law or the log-normal. To achieve this goal we measured citation distribution for a very large and homogeneous data set containing almost all Physics papers published in 1980-1989 and cited by 2008. Surprisingly, we found that (i) the citation distribution is not stationary even 25 years after publication, and (ii) both log-normal and power-law fitting functions fail to describe the extreme tail of the distribution since it develops a runaway behavior.

2 Theoretical framework

In the long time limit the de Solla Price's Cumulative Advantage mechanism [3] generates the following citation distribution

$$p(k) = \frac{B(k + w, \gamma)}{B(w, \gamma - 1)}. \quad (1)$$

Here, k is the number of citations, $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the beta-function, Γ is the gamma-function, and w, γ are

Send offprint requests to:

^a e-mail address: golos@cc.huji.ac.il

parameters. The Eq.1 is known as Waring [18,19,20]) or discrete power-law distribution since for $k \gg w$ it asymptotically approaches the power-law dependence, $p(k) \sim (w/k)^\gamma$. The continuous approximation of Eq.1

$$p(k) \approx (\gamma - 1) \frac{w^{\gamma-1}}{(w+k)^\gamma} \quad (2)$$

is known as the Zipf-Mandelbrot or Pareto-type II distribution and was also used in citation analysis [10,20,22].

Several recent works [13,14,15,16] fitted measured citation distribution with the log-normal

$$p(k) = \frac{1}{k\sigma\sqrt{2\pi}} e^{-\frac{(\ln k - \mu)^2}{2\sigma^2}} \quad (3)$$

function that was modified to describe such non-negative and discrete variable as citations. Although the functions represented by Eqs.1,3 are dramatically different, the fits of the citation distributions based on Eq.1 and on the discrete version of Eq.3 are virtually undistinguishable for a wide (but finite!) range of values. Thus, discrimination between the discrete log-normal and the discrete power-law distributions is quite ambiguous [17,23]. This ambiguity is illustrated by Fig. 1 that shows cumulative probability distribution of citations to 353,268 Physics papers. The data lie just in between the log-normal cdf and discrete power-law cdf.

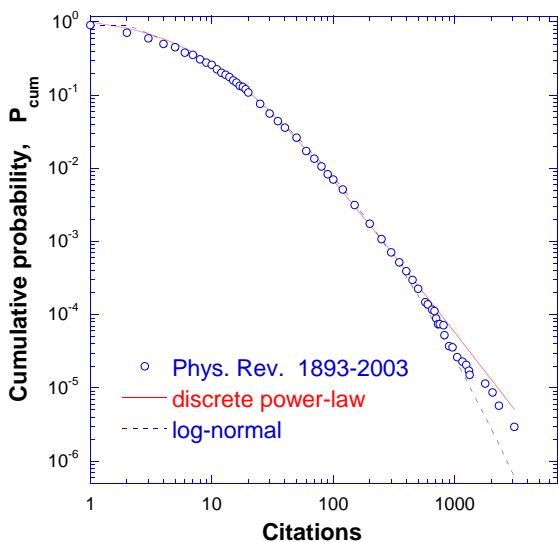


Fig. 1. Cumulative probability distribution (cdf) of citations to 353,268 papers published in Physical Review journals during 1893-2003 and cited by 2003. Only PR to PR citations were counted. The data were adapted from Ref. [13]. The continuous red line shows a fit with the discrete-power-law cdf (Eq.1) with $\gamma = 3.15$, $w = 10.2$. The dashed blue line shows a fit with the log-normal cdf (Eq.3) with $\mu = 1.15$, $\sigma = 1.42$.

The solution of the dilemma "power-law or log-normal" comes from an unexpected direction. Our measurements

demonstrate that (i) the citation distribution follows neither power-law nor log-normal fit, (ii) it is non-stationary, and (iii) the individual values in the tail grow at a much faster rate than the rest of the distribution, indicating the runaway effect. As the sample size and the evolution time increase, the runaways become more prominent. The contribution of non-stationarity in the emergence of distributions with heavy and super-heavy tails has been studied in the context of cities population, wealth distribution and growing networks [24,25,26].

3 Citation Distribution for Physics Papers

We considered research papers in Physics that were published in 82 leading Physical journals in the period from 1980 to 1989 (excluding the overview papers and papers published in the popular science journals) -418,438 papers in total. We measured the number of citations gained by each paper by July 2008. Since the citation life of an ordinary paper rarely exceeds 15 years and the time span between the publication and citation count for the papers in our data set is 20-28 years, we expected that the citation distribution for this data set of "old" papers is stationary.

Figure 2 displays this distribution. The log-normal cdf fits the data only for the papers with less than 400 citations (these constitute 99.7% of all papers), while the tail deviates upward. The discrete power-law cdf fits the data in Fig. 2 much better. However, it also fails to describe the extreme tail of the distribution that contains the papers with more than 1000 citations (0.04 % of all papers). This is most dramatically illustrated in Fig.3 where we plotted the ratio of the measured number of citations to the fitted values using log-normal cdf and discrete power-law cdf. The number of citations for *all* the 190 papers that have more than 1000 citations significantly and increasingly exceed both the power-law and the log-normal fits. It is instructive to compare the onset of this super-heavy or "runaway" tail to the mean number of citations, $m = 24$. Figures 2,3 indicate that the "runaways" become visible at $k/m = 20 - 40$ and extend at least to $k/m = 1000$. The previous smaller scale studies [14,15] were limited to $k/m < 40$ and didn't provide enough statistics to probe the "runaways".

We conclude that citation distribution for the small homogeneous data sets (~ 1000 papers) can be fitted equally well by the log-normal and by the discrete power-law functions. Larger sets ($\sim 10,000$ papers) are better fitted by the discrete power-law cdf. The runaway tail that exceeds even the power-law fit becomes apparent when the number of papers in the set exceeds 10,000. In what follows we focus on the microscopic mechanism responsible for runaways.

4 Microscopic mechanisms of citation dynamics that are able to produce the runaway behavior

1. *Superlinear Yule-Simon process.* The most accepted microscopic mechanism of citation dynamics is the pref-

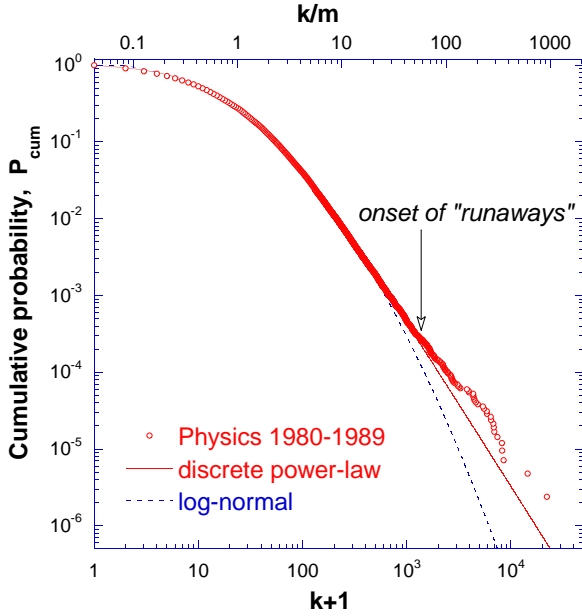


Fig. 2. Cumulative distribution function (cdf) of citations to 418,438 papers published in 82 leading Physical journals in 1980-1989 and cited by July 2008. The horizontal axis is $k+1$ instead of k in order to show uncited papers, $k=0$, on the log-scale. The dashed blue line shows a fit with the log-normal cdf (Eq.3) with $\mu = 2.16$ and $\sigma = 1.38$. The deviation from the fit appears already at $k > 400$. The continuous red line shows a fit with the discrete-power-law cdf (Eq. 1) with $\gamma = 3.15$ and $w = 27.5$. This function fits the data up to $k = 1000$. However, the tail of the distribution with $k > 1000$ deviates upwards even from this power-law fit, indicating "runaway" papers.

erential attachment also known as the cumulative advantage, Gibrat, or Yule process [3, 27, 28, 29]. This mechanism assumes that the citation process is a memoryless Markov chain whereas the citation rate of a paper is determined by the total number of accumulated citations k and the time after publication t ,

$$\frac{dk}{dt} = A(t)(k + k_0)^\alpha \quad (4)$$

Here, A is the aging parameter, α is the attachment exponent, and k_0 is the initial attractiveness. The Refs. [28, 29] showed that the linear preferential attachment, $\alpha = 1$ yields the discrete power-law citation distribution (Eq.1) while the superlinear preferential attachment, $\alpha > 1$, results in the runaway behavior where a few papers contain a finite fraction of all citations.

2. *Stochasticity.* The heavy tail distribution can result from the stochastic and non-stationary character of the dynamic variable [24, 25, 30]. For dynamic variable with lower bound, such as citations, the stochastic mechanism yields the power-law distribution whose exponent is the ratio between the average growth rate and its individual fluctuations [25, 26]. The finite size effects may affect the power-law exponent to the ex-

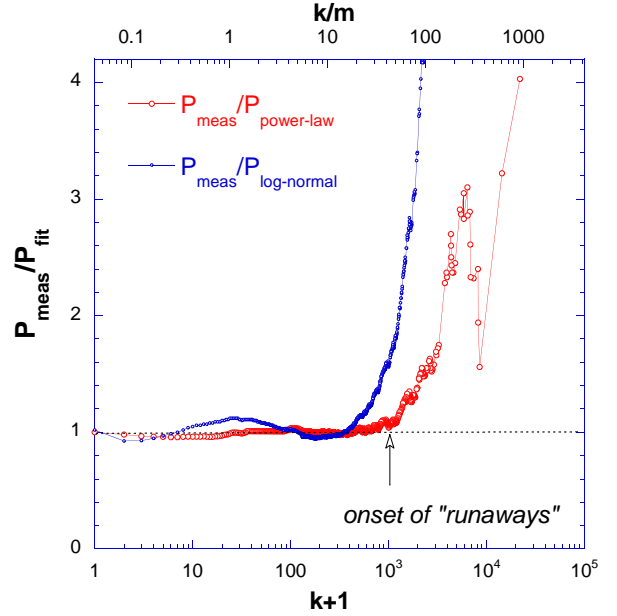


Fig. 3. The ratio of the measured cumulative probability of citations (Fig.2) to the discrete power-law cdf (red circles) and to the log-normal cdf (blue circles). The log-normal fit fails already for $k > 400$. The discrete power-law provides a good fit for $k < 1000$ while pronounced deviation appears above $k > 1000$.

tent of concentrating the entire distribution in just a few individual entries in the end of the heavy-tail [24, 31]. Hence, the stochasticity can partially explain the power-law citation distribution and runaway behavior.

3. *Memory.* If the assumption of a memoryless Markov chain that stands behind Eq.4 is lifted (in other words, citation process has some memory) this can result in the distribution with the runaway tail. Even a weak memory of the past dynamics and its feedback on the rate of the multiplicative random walk may lead to divergent statistical distribution and runaways [32].

To verify to which these mechanisms shape the citation distribution we measured citation dynamics of individual papers. We expected that once the microscopic dynamics is uncovered, we will be able to explain the "runaway tail".

4.1 Microscopic measurements of citation dynamics of individual papers

In distinction to the measurements shown in Fig.2 where we measured aggregated citation distribution for the 418,438 Physics papers published during a decade (1980-1989) and didn't trace citation dynamics of individual papers; here we considered the Physics papers published in *one* year (1984) and analyzed citation history of *each* paper up to July 2008. This data set contains a smaller number of entries (40,195) but it is age-homogeneous.

4.1.1 Divergence of citation dynamics of similar papers

As an example, from all Physics papers published in 1984 we chose the subset containing all those papers that garnered 30-31 citations by the end of 1986. Figure 4 displays citation dynamics of these 89 papers. Although they represent an extremely homogeneous subset (the same field, the same publication year, the same citation prehistory), the divergence in their citation behavior is striking.

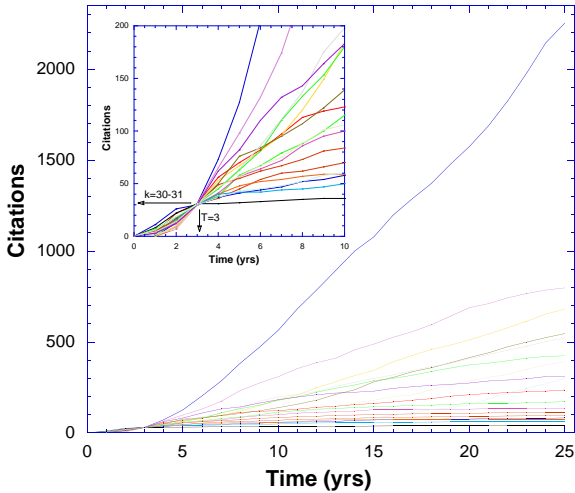


Fig. 4. Citation dynamics of 89 Physics papers published in 1984. We chose all those papers that by 1986 (three years after publication) had 30 or 31 citations. Although the initial citation dynamics of these papers is very similar, it quickly diverges in such a way that after 25 years (in 2008) the number of citations varies between 40 and 2254.

We ranked the papers according to the number of citations garnered by 2008. The list opens with the pair of papers: "Lower critical dimension of the random-field Ising model- a Monte-Carlo study" by D. Andelman, H. Orland, and L.C.R. Wijwardhana, Phys.Rev. Lett. (40 citations); and "Fractional quantum Hall-effect at filling factors up to $\nu = 3$ " by G. Ebert, K. von Klitzing, and J.C.Maan, J. of Physics C- Solid State (49 citations). The lists ends with the papers: "Dynamics of supercooled liquids and the glass transition" by U. Bengtzelius, W. Gotze and A. Sjolander, J. of Physics C- Solid State (798 citations); and "Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals" by M.S. Daw and M.I. Baskes, Phys. Rev. B (2254 citations). All four papers were published actually in the same journals in the same year and they have the same citation pre-history. The two former papers are important works with typical citation dynamics. The two latter papers exhibit strongly different citation dynamics and the last one is a clear runaway. S.Redner [13] already demonstrated that the citation history of the citations classics is strongly individualized. Our analysis substantiates this observation and extends it to all papers.

4.1.2 Analysis of the citation dynamics

We analyzed citation dynamics of all 40,195 Physics papers published in 1984 using the framework of Eq.4. We found $\alpha = 1.2 - 1.28$, $k_0 = 1.1$ and $A = 3.5/(t + 0.3)^2$. Since the deviation from the linearity is small, $\alpha - 1 \ll 1$, and the aging parameter A strongly decays with time, the runaway behavior generated by Eq.4 is too slow. We conclude that while the nonlinear preferential attachment is the dominant mechanism that shapes the observed power-law citation distribution, it can not produce runaways.

As is clearly seen from the Fig.4, the fluctuations of the citation rate of individual papers (the ripple on the continuous lines) is small. Therefore, while the stochasticity affects the shape of the citation distribution its contribution is not dominant. Neither it can be responsible for runaways.

The strong systematic differences between the citation behavior of similar papers shown in Fig.4 is at odds with the preferential attachment paradigm that assumes similar citation dynamics for the papers having the same past number of citations. As borne out by the data, there are strong correlations between the citation rates of the same paper in different years, i.e. the citation process has memory [33]. As we show below, the memory dominates the dynamics of heavily cited papers and turns the citation process in a *de facto* deterministic one, at least for the runaway papers. Indeed, Fig. 4 shows that the number of citations for the most part of the papers comes to saturation after 15 years. However, some papers continue to be cited with undiminished rate even after 25 years. It is clearly seen that the citation behavior of these runaway papers is not erratic, the citation rates in subsequent years are strongly correlated.

4.1.3 Autocorrelation

To characterize the correlations we chose the sets of similar papers and measured autocorrelation between the citation rates of subsequent years. To this end we considered the sets of papers that were published in the same year and that garnered the same number of citations, k , after t years. For each paper i we found the number of citations acquired during two subsequent years - $\Delta k_i(t)$ and $\Delta k_i(t-1)$ - and calculated the Pearson autocorrelation coefficient,

$$c_{t,t-1} = \frac{\left(\overline{\Delta k_i(t) - \Delta k_i(t)} \right) \left(\overline{\Delta k_i(t-1) - \Delta k_i(t-1)} \right)}{\sigma_t \sigma_{t-1}} \quad (5)$$

Here σ_t, σ_{t-1} are standard deviations of the $\Delta k_i(t)$ and $\Delta k_i(t-1)$ distributions, respectively; and the averaging is performed over all papers in the set. Figure 5 shows that for moderately cited papers the autocorrelation is weak, $c_{t,t-1} \ll 1$, as expected for a memoryless process. However, the papers with more than 1000 citations have $c > 0.9$. This means that their citation behavior is almost deterministic. While the citation rate of the most part of

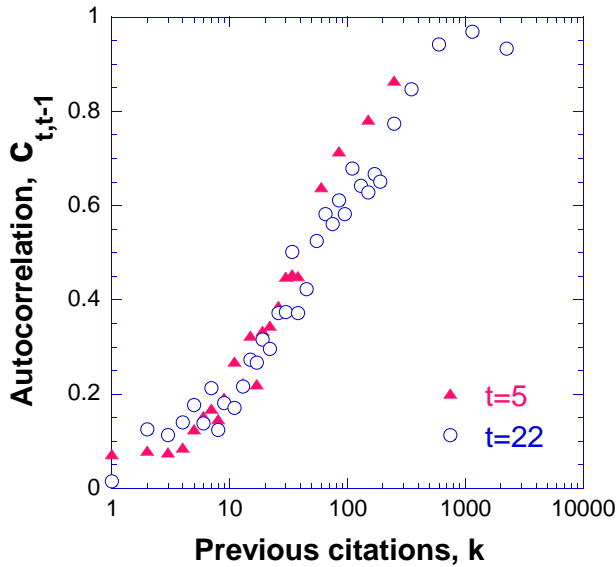


Fig. 5. The Pearson autocorrelation coefficient for additional citations (Eq.5). t is the number of years after publication. $c_{t,t-1}$ steadily increases with the number of previous citations k , and exceeds 0.9 for the papers with more than 1000 citations.

the papers decays with time, these papers continue to be highly-cited and eventually develop into runaways. (Indeed, the onset of the runaway tail in Fig.3 occurs at the same number of citations, $k = 1000$, at which the autocorrelation coefficient approaches unity.) In the long run, the citation rate of these runaways shall nevertheless decrease due to yet another mechanism - these prominent papers eventually become common knowledge and need not to be cited anymore.

4.1.4 Numerical simulations

Our further studies of citation statistics (not shown here) combined correlation measurements with other parameters of the correlated Poisson process governing the citation dynamics of each paper, namely, stochastic and deterministic parts of the Eq.4. We performed numerical simulations of citation dynamics based on these measurements and without any additional parameters, and found aggregated citation distribution. The simulations developed a runaway population of papers consistent with that observed in Fig.2. The simulations based only on the stochastic version of the Eq.4 (without correlations) yielded discrete power-law citation distribution and no runaways.

5 Discussion

Unlike other fields where "dragon king" events have negative emotional connotation and are associated with catastrophes and crashes, the runaways in citation statistics

have positive connotation and indicate very important papers in the field. The social and epistemological implications of our study might contribute to the current effort of evaluating and predicting scientific production in terms of citations.

Our results are important also in a broader context of the dynamics of complex systems. Our citation statistics results add themselves to the relatively small number of systems for which the microscopic elementary laws were uncovered and whose macroscopic consequences (obtained theoretically and by simulation) passed successfully the confrontation with the empirically observed macroscopic phenomena [34,35,36].

The present study should be viewed as a part of the more generic effort of understanding autocatalytic dynamics as the bridge that allows the promotion of simple microscopic interactions to complex, macroscopic collective phenomena.

6 Conclusions

- 1 Aggregate citation distribution for the set of "old" Physics papers is better fitted by the discrete power-law (Waring) distribution rather than by the log-normal distribution.
- 2 Both power-law and log-normal fits severely underestimate the extreme tail of the citation distribution that contain papers with more than few thousands citations ($\sim 0.05\%$ of all papers). These "runaways" exhibit significant autocorrelation in their citation rate (the Pearson autocorrelation coefficient exceeds 0.9.) in such a way that it becomes deterministic.
- 3 In the long run, citation distribution for the age-and field-homogeneous set of papers reveals two populations: ordinary papers whose citations come to saturation after 15-20 years and "runaways" that continue to be cited even after 20 years.

7 Acknowledgments

We are grateful to B. Meerson, G.J. Peterson, S. Redner, P. Richmond, N. Shnerb, and J. Weissbuch for valuable discussions. We appreciate important correspondence with Q. Burrell, S. Fortunato, and J.C. Phillips. This work was partly supported by the CO3, Daphnet.

References

1. Newman MEJ (2005) Power laws, Pareto distributions and Zipfs law. *Contemporary Physics* 46:323-351.
2. de Solla Price D (1965) Networks of scientific papers. *Science* 149:510-515.

3. de Solla Price D (1976) A General Theory of Bibliometric and Other Cumulative Advantage Processes. *J Am Soc for Inform Science* 27: 292-306.
4. Redner S (1998) How popular is your paper? An empirical study of the citation distribution. *Eur Phys J B* 4: 131-134.
5. Lehmann S, Lautrup B, Jackson AD (2003) Citation networks in high energy physics. *Phys Rev E* 68:026113.
6. Borner K, Maru JT, Goldstone RL (2004) The simultaneous evolution of author and paper networks. *Proc Natl Acad Sci USA* 101:5266-5273.
7. van Raan AFJ (2005) Reference-based publication networks with episodic memories. *Scientometrics* 63:549-566.
8. Newman MEJ (2009) The first-mover advantage in scientific publication. *European Physics Letters*, 86:68001-p1-p6.
9. Peterson GJ, Presse S, Dill KA (2010) Nonuniversal power law scaling in the probability distribution of scientific citations *Proc. Natl. Acad. Sci. USA* 107:16023-16027.
10. Wallace ML, Lariviere V, Gingras Y (2009) Modeling a century of citation distributions. *Journal of Informetrics* 3: 296-303.
11. Phillips JC (2009) Topology and the Web of Twentieth Century Science, <http://arxiv.org/abs/0909.5363v1>.
12. Laherrere J, and Sornette D (1998) Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales, *Eur. Phys. J. B* 2:525-539.
13. Redner S (2005) Citation Statistics from 110 Years of Physical Review, *Physics Today* 58:49-54.
14. Stringer MJ, Sales-Pardo M, Amaral LAN (2008) Effectiveness of Journal Ranking Schemes as a Tool for Locating Information. *PLoS ONE* 3(2): e1683.
15. Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: toward an objective measure of scientific impact. *Proc. Natl. Acad. Sci. USA* 105:17268-17272.
16. Petersen AM, Wang F, Stanley HE (2010) Methods for measuring the citations and productivity of scientists across time and discipline, *Physical Review E* 81:036114.
17. Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. *SIAM Rev.* 51:661-703.
18. Burrell QL (2005) The use of the generalized Waring process in modelling informetric data. *Scientometrics* 64:247-270.
19. Mingers J, Burrell QL (2006) Modeling citation behavior in Management Science journals. *Information Processing and Management* 42:1451 -1464.
20. Glanzel W (2007) Characteristic scores and scales. A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics* 1:92-102.
21. van Raan AFG (2001) Two-step competition process leads to quasi power-law income distributions Application to scientific publication and citation distributions. *Physica A* 298:530536.
22. Tsallis C, de Albuquerque MP (2000) Are citations of scientific papers a case of nonextensivity? *European Journal of Physics B* 13:777-780.
23. Egghe L, Rousseau R (2003) Size-frequency and rank-frequency relations, power laws and exponential relations: a unified approach, *Progress in Natural Science* 13:478-480.
24. Blank A, Solomon S (2000), Power laws in cities population, financial markets and internet sites (scaling in systems with a variable number of components), *Physica A* 287:279-288.
25. Malcai O, Biham O, Solomon S (1999), Power-law distributions and Levy-stable intermittent fluctuations in stochastic systems of many autocatalytic elements, *Phys. Rev. E* 60:1299-1303.
26. Klass OS, Biham O, Levy M, Malcai O, Solomon S (2007) The Forbes 400, the Pareto power-law and efficient markets. *Eur Phys J B* 55:143-147.
27. Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74:47-97.
28. Krapivsky PL, Redner S, Leyvraz F (2000) Connectivity of growing random networks *Phys. Rev. Lett.* 85:4629-4362.
29. Dorogovtsev SN, Mendes JFF, Samukhin AN (2000) Structure of Growing Networks with Preferential Linking. *Physical Review Letters* 85:4633-4636.
30. Levy M, Solomon S (1996), Dynamical Explanation for the Emergence of Power Law in a Stock Market Model, *International Journal of Modern Physics C* 7:65-72.
31. Huang ZF, Solomon S (2001) Finite market size as a source of extreme wealth inequality and market instability, *Physica A* 294:503-513.
32. Louzoun Y, Solomon S (2001) Volatility driven markets in a generalized Lotka-Volterra formalism, *Physica A* 302:220-233.
33. Cattuto C, Loreto V, Servedio VDP (2006) A Yule-Simon process with memory, *Europhysics Letters* 76:208-214.
34. Yaari G, Solomon S, Rakocy K, Nowak A (2008), Microscopic Study Reveals the Singular Origins of Growth, *European Physics Journal B* 62:505-513.
35. Challet D, Solomon S, Yaari G (2009) The Universal Shape of Economic Recession and Recovery after a Shock, *Economics* 3(36):1-24.
36. Dover Y, Moulet S, Solomon S, Yaari G (2010) Do all economies grow equally fast? *Risk and Decisions Analysis*; 1(3):171-185.

